ELTE | IK

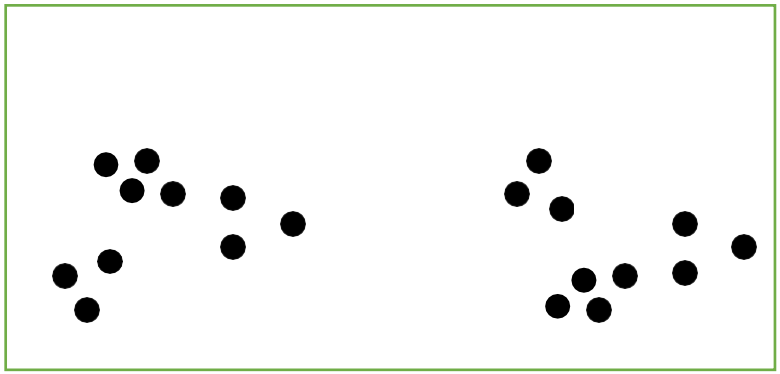
INFORMATIKAI KAR

Adattárházak, adatbányászati technológiák GY.

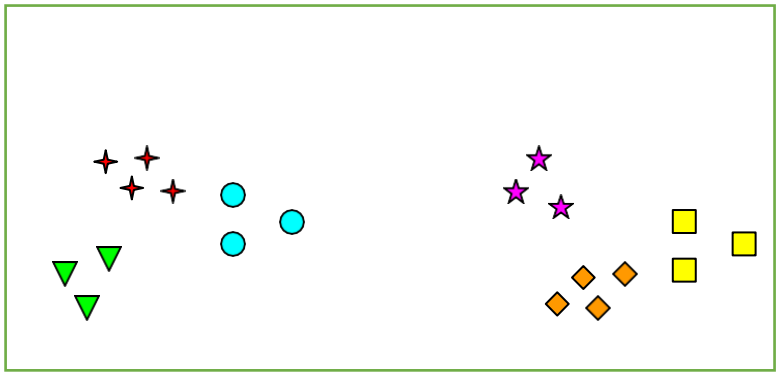
Gépi tanulás kategóriái

- Felügyelt tanulás (supervised learning)
- Nem felügyelt tanulás (unsupervised leraning)
- Megerősítéses tanulás (reinforcement learning)

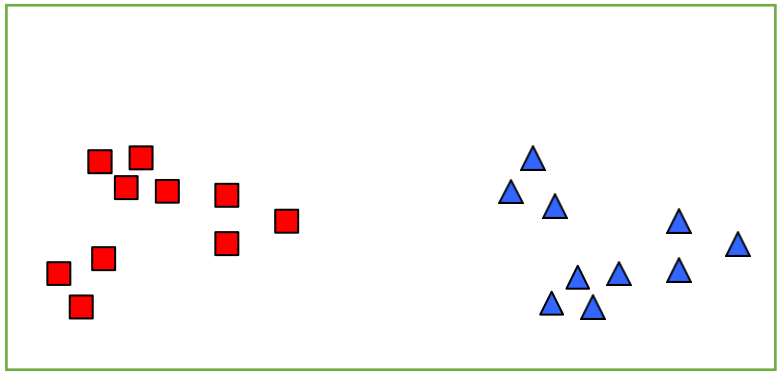
Hány csoport van az adathalmazban?



Hány klaszter?



Hat klaszter

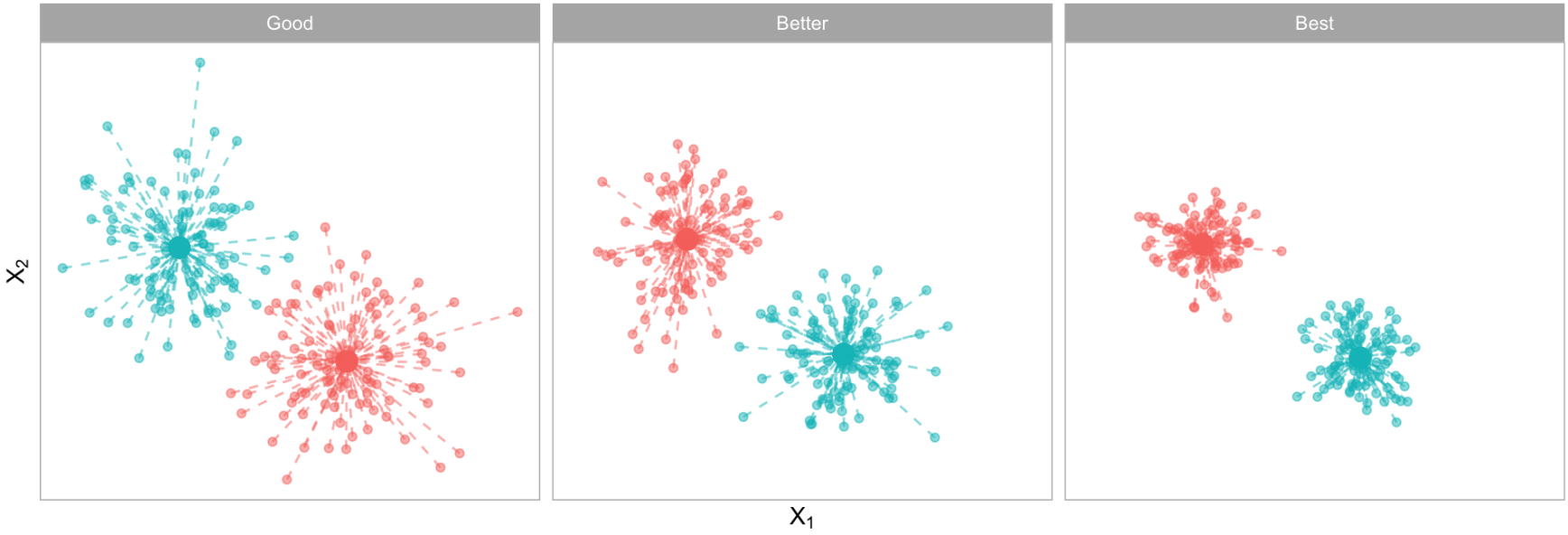


Két klaszter

Négy klaszter

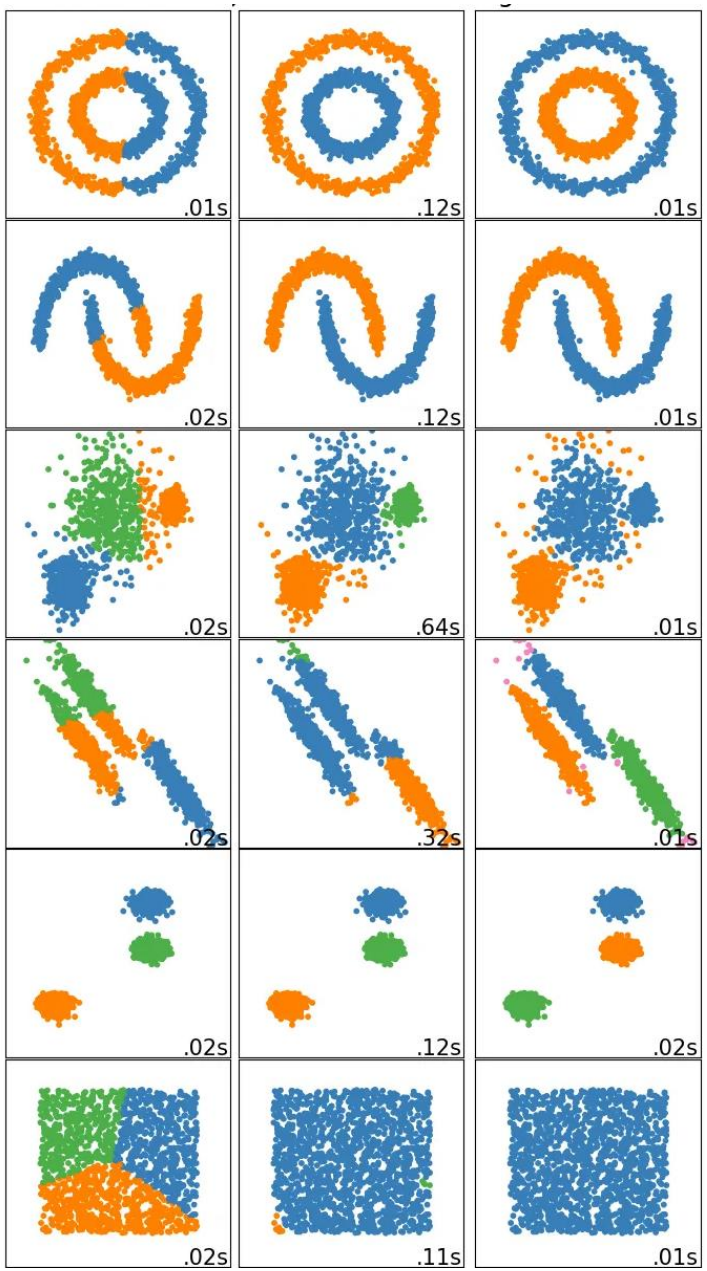
Klaszterezés kiértékelése

- Withiness – egy klaszteren belüli pontok klaszterközépponthoz vett távolságaik négyzetének az összege
- Total withiness – minden klaszter withiness értékének az összege



Típusok

- K-Means
- Hierarchikus
- DBSCAN



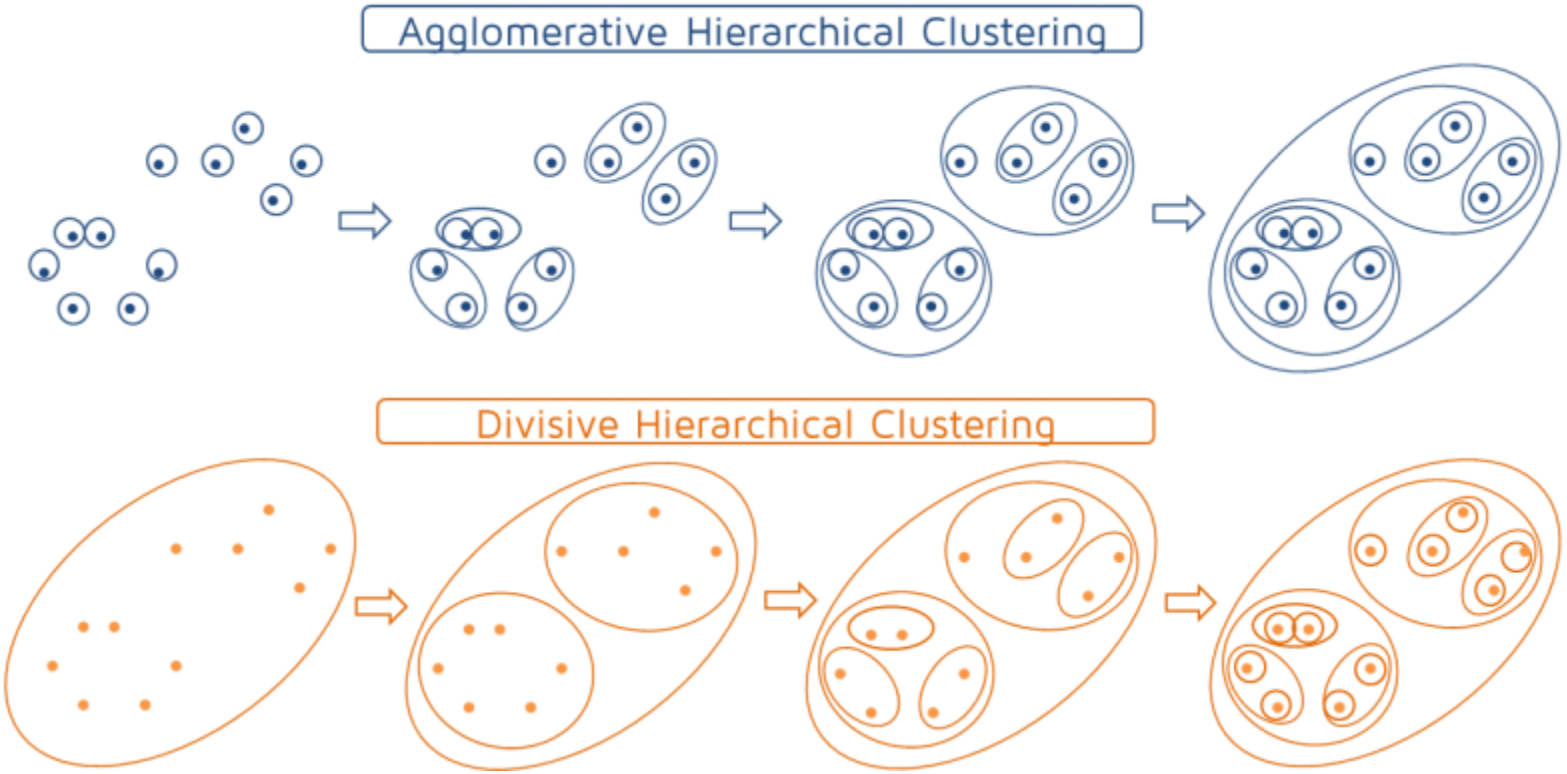
Vizualizáció

- K-Means:

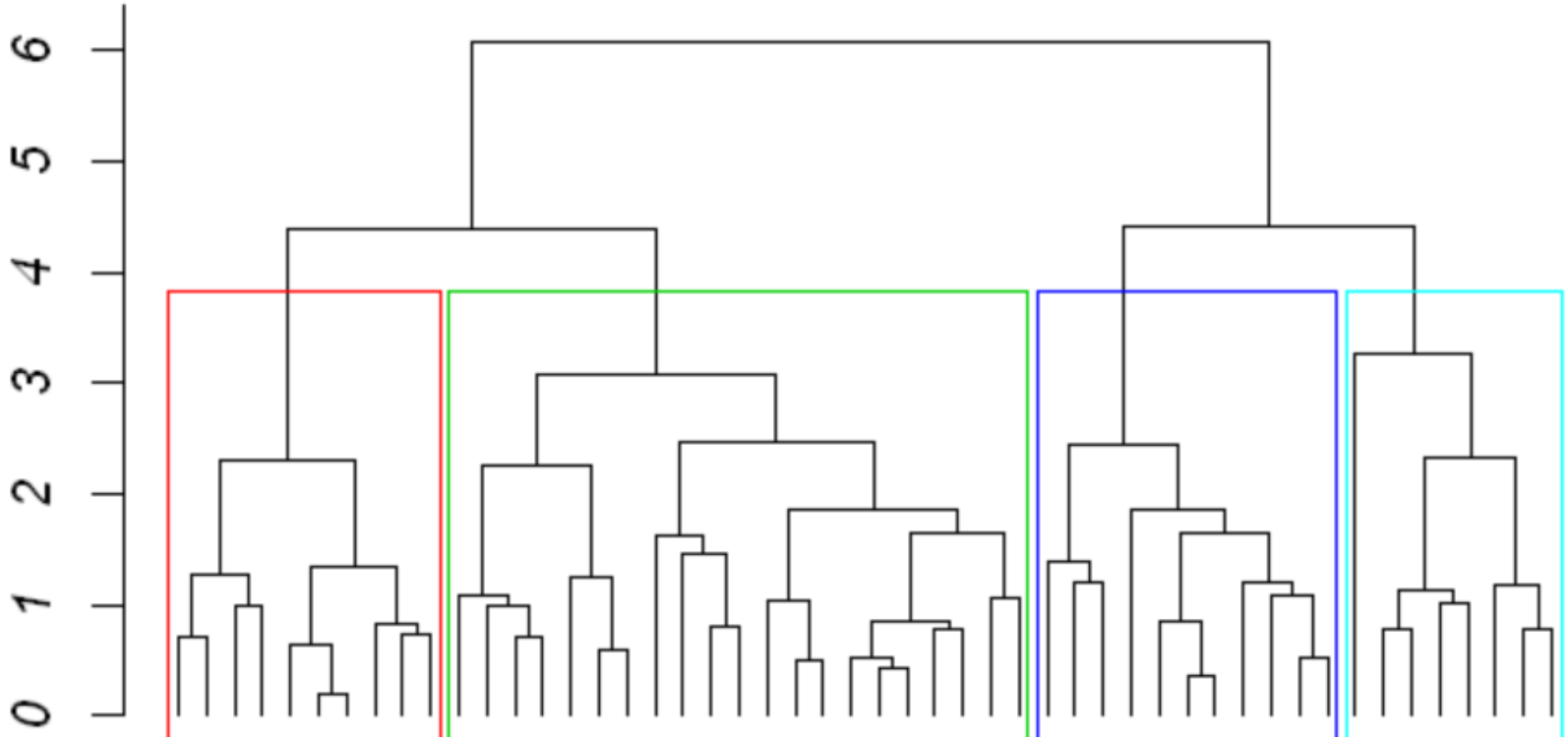
  https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
- DBSCAN:

  https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Hierarchikus klaszterezés

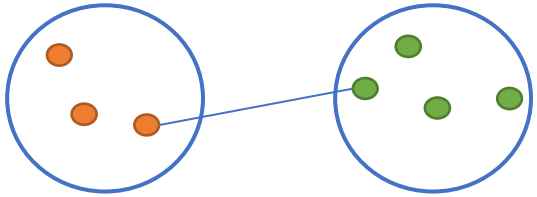


Dendogram

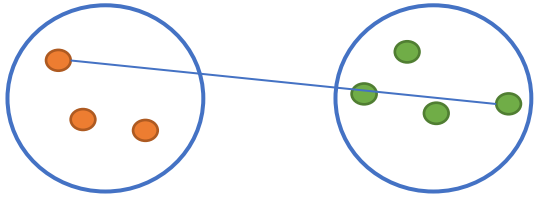


Klaszter távolság metrikák

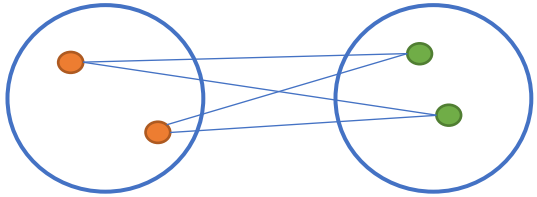
- Single linkage – a klaszterek egymáshoz legközelebb lévő pontjai közötti távolság



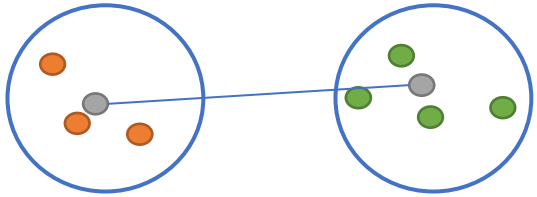
- Complete linkage – a klaszterek egymáshoz legtávolabb lévő pontjai közötti távolság



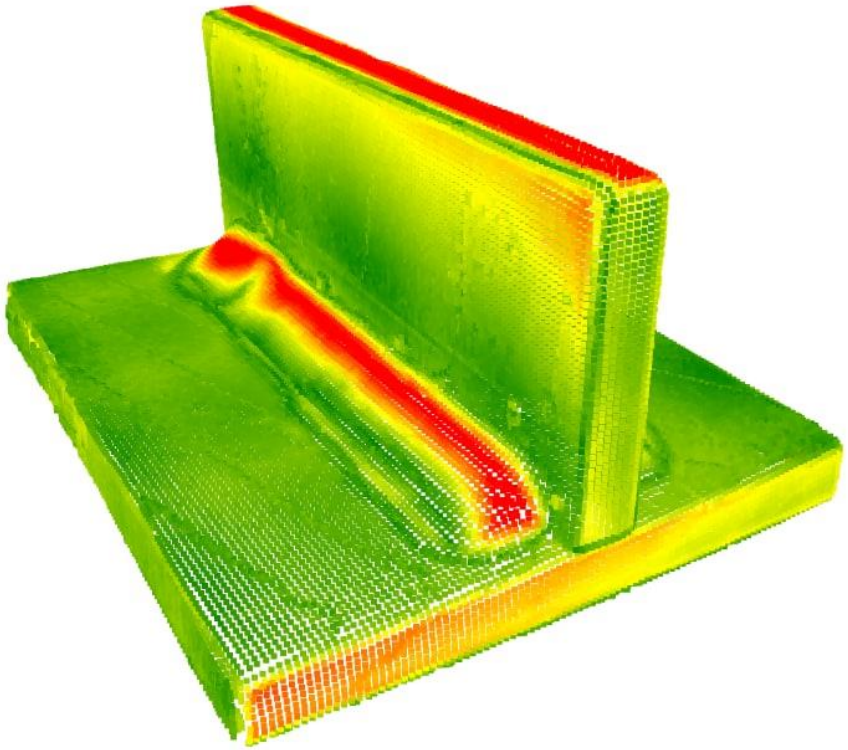
- Average linkage – a klaszterek közötti minden pontpár távolságának az átlaga

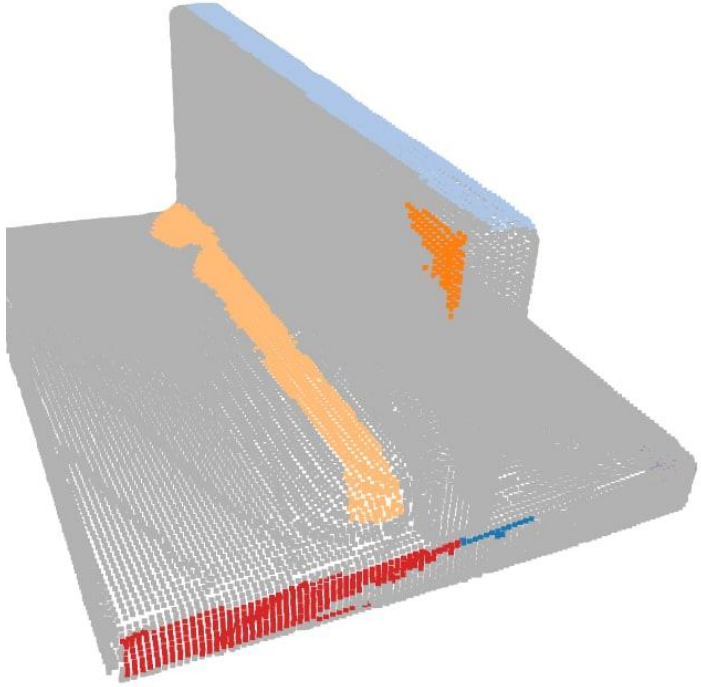


- Centroid módszer – a klaszterek centroidja közötti távolság

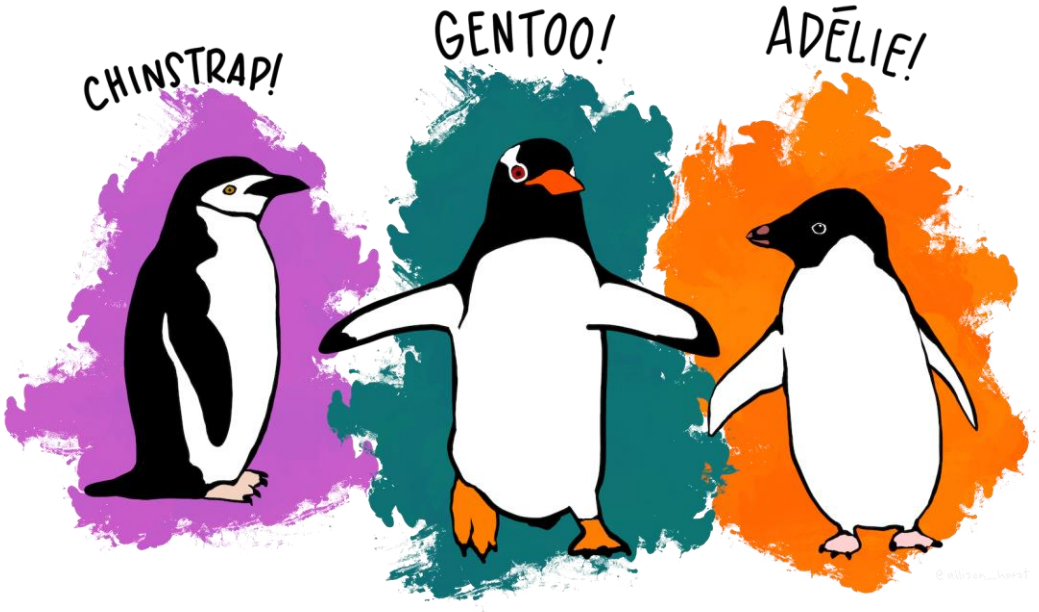


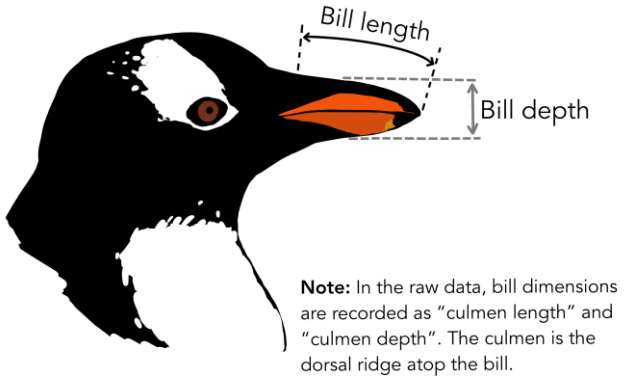
Sűrűség alapú klaszterezés példa





Adathalmaz: palmerpenguins





Note: In the raw data, bill dimensions are recorded as "culmen length" and "culmen depth". The culmen is the dorsal ridge atop the bill.

ELTE | IK